Comparative Study on Sentiment Analysis of Human Speech using DNN and CNN

Sayak Ghosal, Saumya Roy, and Rituparna Basak

Abstract Communication is a way of exchanging thoughts through emotions. In this paper, we have proposed a method where human speech is converted into digital input. The digitized sound is then fed into the proposed models, and the voice of every person is classified into discrete emotional characteristics by its pitch, intensity, timbre, speech rate, and pauses. In the proposed method, we have drawn a comparative study between sentiment analysis of human speech using deep convolutional neural network (CNN) and dense deep neural networks (DNNs). In this method, multiscale area attention is applied in deep CNN as well as dense DNN to obtain emotional characteristics with wide range of granularities and therefore, the classifier can predict a wide range of emotions on a broad scale classification.

Keywords Sentiment analysis · Audio analysis · Deep learning · Neural networks · Emotion detection · Deep neural network (DNN) · Convolutional neural network (CNN)

1 Introduction

Speech is considered to be the most valuable and widely used means of communication. Speech emotion recognition (SER) has wide application-perspectives on psychological assessment, robotics, etc. For example, a doctor treating a patient suffering from depression can keep a track of his patient's development and design a recovery plan according to the patient's speech sentiments. Over the past few years, there has been a consequential development in the field of analyzing the emotions in speech with deep learning. However, deficiencies still exist in SER research due to lack of substantial model accuracy, deficit of useful dataset, and lack of computing resources. In SER, emotion may depict distinct energy patterns in spectrograms with varied granularity of areas. The change in different types of emotions brings changes in energy patterns in spectrograms. Typically, in SER attention, models are

S. Ghosal · S. Roy · R. Basak (✉)
University of Engineering and Management, Kolkata 700156, India
e-mail: rituparna.basak@uem.edu.in

J. K. Mandal et al. (eds.), *Topical Drifts in Intelligent Computing*, Lecture Notes in Networks and Systems 426, https://doi.org/10.1007/978-981-19-0745-6_10

commonly optimized on a fixed scale that may confine the model's capability to deal with diverse areas and granularities. An attention neural network [1] classifier of SER is usually optimized on a fixed attention granularity. In SER, distinct energy patterns are depicted in spectrograms which have varied granularity of areas.

In the proposed method, this constraint is removed by applying multiscale area attention in deep CNN [2] as well as dense DNN to obtain emotional characteristics with a wide range of granularities and therefore, the classifier can predict a wide range of emotions on a broad scale of classification. For example, sentiments such as annoyance and joy have different levels of intensity, so instead of just categorizing the presence or absence of the emotion, level of intensity of the emotion can be identified as well. Hence, a comparative study using both the models is conducted.

Models that have been used for SER previously suffer a problem of sample scarcity. A novel approach has been used to deal with data scarcity. Augmentation of data with addition of stretching, pitch modification, and noise insertion has been used for this reason. This adds variation to the dataset, and it also improves the chances of getting better accuracy. For example, a training data with all anger emotions expressed in higher pitch will now have the same outcome as with an emotion in lower pitch, hence maximizing the chances of identifying anger, when spoken in a lower timbre. Similarly, the sample set might consist of audio of fast speakers and be essentially biased. Stretching helps in removing this bias. Adding noise to the data is proven to be a useful tool for classifying real-time date. To the effectiveness of the proposed method, extensive experiments are carried out on RAVDESS [3], CREMA-D [4], TESS-D [5] dataset.

2 Literature Survey

In the paper [6], Fayek et al. propose a real-time speech emotion recognition system based on end-to-end (E2E) learning. From a one second frame of raw speech spectrograms, the technique of deep neural network is used to study the emotions. A deep hierarchical framework, pragmatic optimization, and data augmentation help in achieving the desired results. Promising results are reported.

In the paper [7], well organized procedure has been provided by author, for implementing SER political debates. The emphasis is laid on manufacturing the outcome and then to prepare visualization of the said results. Two alternative approaches have been considered, such as a classification-oriented approach and a lexicon-oriented approach. In the former universal and domain-oriented sentiment, lexicons are used. Two general methods for implementing domain-oriented lexicons-based approach have also been considered. These are (a) direct generation and (b) adaptation. Direct generation focuses on producing exclusive lexicons depending upon the data labels. Adaptation considers a common and inclusive lexicon-based approach and adjusting it as per necessity to develop it into a non-generic and exclusive symbol of a particular domain. The results obtained from the above discussed approaches were considered and compared with the "classification-based" approach. By observing and analyzing the attitude of the political speakers in the debates, the sentiment mining approaches were compared. Collective labeled speech data were considered, which were of political significance which was extracted from debating transcripts. The outcome of the comparison helped them realize that using sentiment mining, the speakers attitude can be determined conclusively. The proposed debate graph extraction (DGE) framework, in its functioning, effectively extracts the debate graphs from political debate transcripts. They proposed to graphically represent debates with speakers as nodes. In this framework, the speakers are represented as nodes, with nodes having specific labels and links between nodes. These links depend upon the exchange of speeches. The labels on the nodes depended upon the sentiment of the speakers. The attitude of the speaker was then used to classify a link as supporting or non-supporting. If the outcome of both speakers was same, i.e., both positive or both negative, then the link was categorized as supporting or else it was categorized as opposing. Visualization of results was carried out via graphs that represent the essence of the debate, in an abstract manner.

In the paper [8], Tripathi et al. proposed automatic sentiment detection system for natural audio streams. Part of speech tagging and maximum entropy modeling (ME) has been used as the suggested technique to develop a sentiment detection model that was text-based in nature. The number of model boundaries in ME was reduced drastically by an attuning technique while conserving the classification capability. Using decoded automatic speech recognition (ASR) transcripts and the ME sentiment model, sentiments of YouTube videos were able to be determined. As evaluation, they have gathered motivating classification accuracy. According to the results, analysis showed that performance on sentiment analysis on spontaneous speech data is possible in spite of word error rates.

3 Problem Statement

The human speech is the most innate way of expressing oneself. We know emotions play an important role in communication analysis, and the detection of the same is significantly important in today's digital world of remote communication. In text-based classification certain emotions like sarcasm, dual meaning sentences cannot be identified. Tonal qualities of the voice are required to classify the emotions more accurately. An SER system can thus be defined as a collection of methodologies that classifies speech signals to detect embedded emotions. The human speech contains many features different to each individual. If we consider all those features while training the model, then the model will be biased to the training set which is not desired. So, we have considered only the properties common to human voices like loudness, timbre, and quality. Our attempt lies in trying to detect underlying emotions embedded in speech through analysis of the acoustic features of the audio recording.

4 Dataset

Our dataset focusses on RAVDESS, CREMA-D, TESS-D dataset. These three instances of audio datasets used during our analysis and contain the vocal emotional expressions of sentences spoken in a range of varied emotional indexes (joy, grief, rage, agitation, annoyance, and calm). Total of 1440 and 7,440 clips of 115 actors were collected with diverse ethnic background, and it was merged into 8882 files. We have worked only with the audio recordings of the audiovisual data. The sentences are spoken by trained actors belonging to a variety of races and ethnicities (Latino Americans, African, American, Asian, Caucasian). The sentences are classified using one of six different emotions (joy, grief, rage, agitation, annoyance, and calm) and four different emotion levels (low, medium, high, and unspecified). The audio file format used is WAV. To obtain variance in the data, noise is introduced which is stretched and then inserted. Stretching is important for data that have audio cues which are short in nature. To analyze in real-time, noise in an essential component that should not add bias to the results. Therefore, noise is introduced into the training sample.

5 Proposed Solution

Three classes of features can be mainly identified in a speech. These can be classified as lexical features, the visual features, and the acoustic features. For example: the various expressions of the speaker, the terminology used, and properties like vocal quality, pitch, anxiety, noise, energy, etc. (Fig. 1).

Analysis of lingual features requires a script of the speech. However, it will require a processing and extraction of text from speech in order to analyze sentiments from real-time audio. During analyzation of visual features, it requires the access to the video of conversations, and it is not in the scope of this research. Therefore, analysis of the auditory features is done in this work, since analysis of the acoustic features is

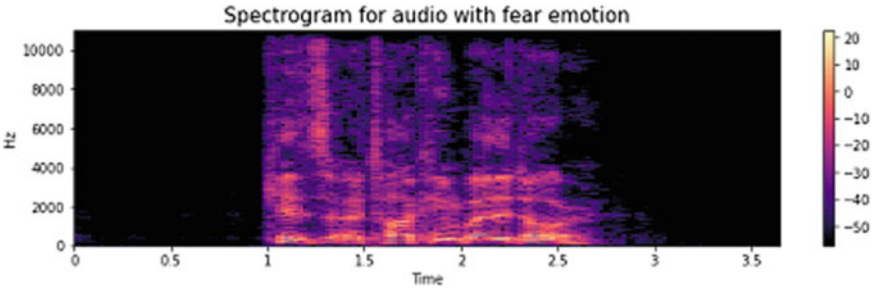


Fig. 1 Spectrogram plotted for audio with fear emotions

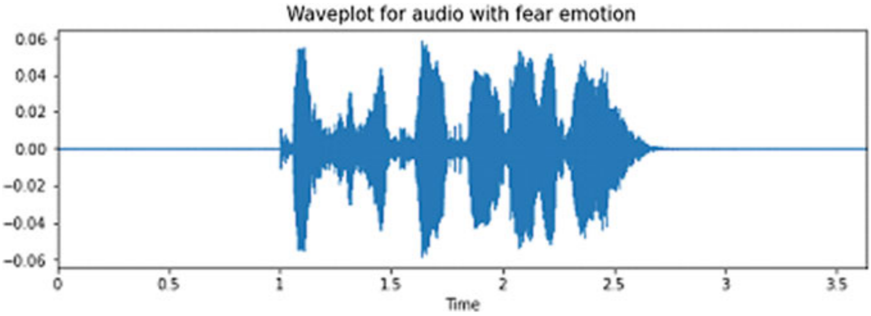


Fig. 2 Various categories of sentiments that our model predicts

possible in real-time. Audio from real-time conversations is extracted and analyzed in order to accomplice the task (Fig. 2).

The representation of emotions can be done in two ways:

- Discrete Classification: Emotions were classified into distinct labels like rage, calmness, neutral, cheerfulness, and joy.
- Dimensional Representation: Representations of emotions with dimensional categories such as activation energy (on a low to high scale), valence (on a negative to positive scale), or/and dominance (on an active to passive scale).

The two mentioned approaches each have their distinct advantages and disadvantages. The dimensional representation approach is an elaborative process, but there is a deficiency of annotated audio data in the dimensional format. The discrete classification is more straightforward and less resource hungry to implement.

In discrete classification approach, emotions are classified on a specified scale for the analysis. Emotions are classified using the trained model and predicted as discrete outputs. This approach is easier to implement and understand and as such has greater outreach, and thus we have focused on this approach.

Dense DNN Architecture:

The different acoustic features from the training sample set are extracted. Features like MFCC, zero crossing rate, chroma STFT, and mel spectrogram have been used to train our model. The MFCCs stand for mel frequency cepstral coefficients. These form cepstral representation for non-linear frequency bands that are distributed according to the mel scale. Zero crossing rate identifies the rate of sign changes of a signal for a particular frame duration. The feature chroma portrays the 12-element representation of the 12 equally tempered pitched classes. These are called the chroma coefficients.

Once the features are extracted, the data are preprocessed and prepared, by noise injection and stretching for audio augmentation. Thereafter, they are fed into the DNN model architecture, for analysis. Our inputs are represented by 128 dimensions, with a dropout of 0.2. We have primarily utilized ReLu activation function to obtain better results, with a final layer having Softmax activation. We have implemented 5 layers.

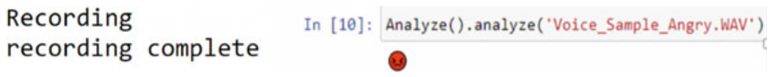
2D CNN Architecture:

The different acoustic features from the training sample set are extracted. Features like MFCC, zero crossing rate, chroma STFT, and mel spectrogram have been used to train our model. The MFCCs stand for mel frequency cepstral coefficients. These form cepstral representation for non-linear frequency bands that are distributed according to the mel scale. Zero crossing rate identifies the rate of sign changes of a signal for a particular frame duration. The feature chroma portrays the 12-element representation of the 12 equally tempered pitched classes. These are called the chroma coefficients. Once the features are extracted, the data are preprocessed and prepared, by noise injection and stretching for audio augmentation. Thereafter, they are fed into our 2D-CNN followed by a dense DNN model architecture, for analysis. Our inputs are represented by 256 dimensions, with a dropout of 0.3. We have primarily utilized ReLu activation function to obtain better results. We have implemented 4 2D-CNN layers and 2 dense layers.

In Fig. 3, we have taken an audiovisual file as an input for testing. The audio file consists of the voice of an actor who portrays the emotion of anger. The algorithm predicted the emoticon for the emotion portrayed correctly.

The proposed comparative study consists of two separate classification models. The first model is constructed primarily on 2D-CNN, connected to a dense DNN. The second approach that is considered is using only dense DNN. After careful observations of the outputs, the realizations of both the models are portrayed below.

Figure 4 shown above plots the output of the training and testing for 2D-CNN model. In the X-axis, the number of epochs is plotted. In Fig. 4a, training and testing



```
Recording
recording complete
```

```
In [10]: Analyze().analyze('Voice_Sample_Angry.WAV')
```

Fig. 3 Analysis of audiovisual data

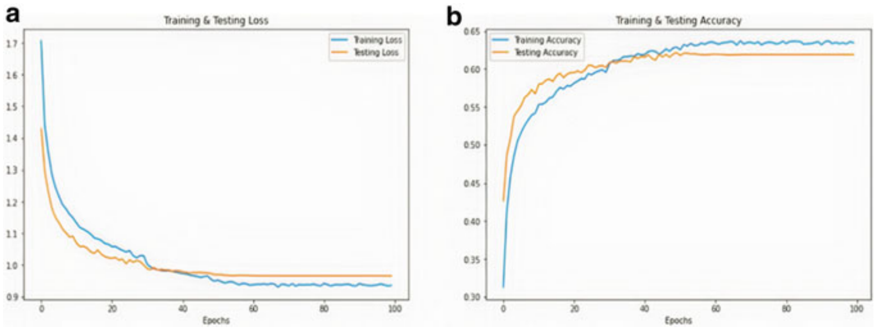


Fig. 4 **a** Training and testing loss 2D-CNN model, **b** training and testing accuracy of 2D-CNN model

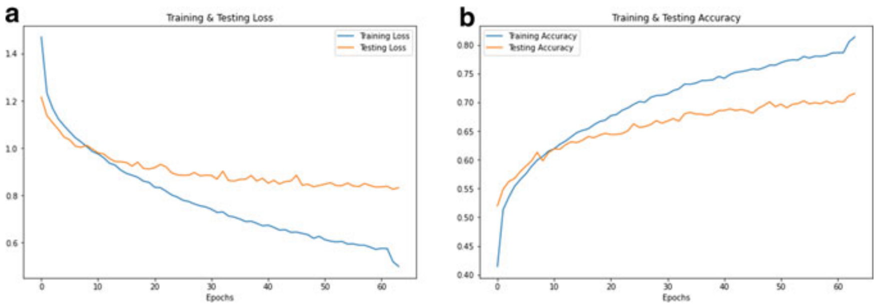


Fig. 5 **a** Training and testing loss dense DNN model, **b** training and testing accuracy of dense DNN model

loss of 2D-CNN model is shown and Fig. 4b training and testing accuracy of 2D-CNN model are shown. On analysis of the graphs, each epoch in both the graphs tend to merge at a point which is desired. After the point of saturation, there is no significant change in the difference between training and testing loss of the model and also training and testing accuracy of the model which shows the strong integrity of the proposed model.

Figure 5 shown above plots the output of the training and testing for the dense DNN model. In the X-axis, number of epochs is plotted.

In Fig. 5a, training and testing loss of dense DNN model are shown and Fig. 5b training and testing accuracy of dense DNN model are shown. On analysis of the graphs, with each epoch in both the graphs, the lines tend to merge at a point. After merging, the lines diverge, which indicates significant variance between the training and the testing loss. With higher number of epochs, the loss decreases since the model refuses to achieve saturation. In case of the accuracy plot, with higher number of epochs, the accuracy increases since the model refuses to achieve saturation.

Experimental Analysis:

The results obtained that of dense DNN model supersedes that of its counterpart, i.e., the model comprising of 2D-CNN connected to dense DNN. The accuracy obtained for the dense DNN model is 71.52% whereas that of its counterpart is 61.89%. The loss incurred in the dense DNN is around 0.83 whereas that of the 2D-CNN model is around 1.0. However, the dense DNN model shows a tendency to overfit on higher epochs, unlike the 2D-CNN model. The advantage of the latter is that it considers lower number of features in each step and operates with a lesser number of coefficients that prevent overfitting.

6 Conclusion

The achieved accuracy in case of dense DNN is of 71.52 and 61.89% in case of 2D-CNN. Though the accuracy of the dense DNN is higher, the model tends to overfit the given sample data. Thus, from the above study, it can be concluded that both the models provide different perspective to the problem of SER and their use is completely dependent on the type of utilization required. The proposed model is able to provide some noteworthy results that can have myriad of applications. Efficient utilization of the audio signals and their tone, pitch, and granularity can help in detection of lies, mimicry, as well as mental state of a person. Furthermore, for exploring broader spectrum like analysis and interpretation, such as analysis of interviews and interrogations, multimodal sentiment analysis should be taken into considerations.

References

1. Yin W, Schütze H, Xiang B, Zhou B (2016) ABCNN: attention-based convolutional neural network for modeling sentence Pairs. In: Transactions of the association for computational linguistics, vol 4. pp 259–272. https://doi.org/10.1162/tacl_a_00097
2. Goecke R, Potamianos G, Neti C (2002) Noisy audio feature enhancement using audio-visual speech data. In: 2002 IEEE international conference on acoustics, speech, and signal processing. pp II-2025–II-2028. https://doi.org/10.1109/ICASSP.2002.5745030
3. Kurpukdee N, Kasuriya S, Chunwijitra V, Wutiwiwatchai C, Lamsrichan P (2017) A study of support vector machines for emotional speech recognition. In: 2017 8th international conference of information and communication technology for embedded systems (IC-ICTES). pp 1–6. https://doi.org/10.1109/ICTEmSys.2017.7958773
4. Abburi H (June 2017) Audio and text based multimodal sentiment analysis using features extracted from selective regions and deep neural networks. International institute of information technology Hyderabad–500032, India
5. Iqbal M. MFCC and machine learning based speech emotion recognition over TESS and IEMOCAP datasets. https://doi.org/10.33897/fujeas.v1i2.32
6. Salah Z (May 2014) Machine learning and sentiment analysis approaches for the analysis of parliamentary debates. ISNI:0000-0004-5365-4219
7. Tripathi S, Kumar A, Ramesh A, Singh C, Yenigalla P. Deep learning based emotion recignition system using peech features and transcription. arXiv:1906.05681 [eess.AS]
8. Lugović S, Dunđer I, Horvat M (30 May–3 June, 2016) Techniques and applications of emotion recognition in speech. In: MIPRO. Opatija, Croat
9. Xu M, Zhang F, Zhang W (2021) Head Fusion: improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. In: IEEE Access, vol 9. pp 74539–74549, 202. https://doi.org/10.1109/ACCESS.2021.3067460
10. Tai O, Liu Y, Huang J, Yu X, Aljbawi B (1 May 2021) Neural attention frameworks for explainable recommendation. IEEE Trans Knowl Data Eng 33(5):2137–2150. https://doi.org/10.1109/TKDE.2019.2953157
11. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R (1 Oct–Dec 2014) CREMA-D: crowd-sourced emotional multimodal actors dataset. IEEE Trans Affect Comput 5(4):377–390. https://doi.org/10.1109/TAFFC.2014.2336244
12. Zheng F, Zhang G, Song Z. Comparison of different implementations of MFCC. J Comput Sci Technol 16:582–589. https://doi.org/10.1007/BF02943243

13. Hermansky H, Cox LA (1991) Perceptual linear predictive (Plp) analysis-rsynthesis technique. In: Final program and paper summaries 1991 IEEE ASSP workshop on applications of signal processing to audio and acoustics. pp. 0_37–0_38. https://doi.org/10.1109/ASPAA.1991.634094
14. Cheong S, Oh SH, Lee S-Y (March 2004) Support vector machines with binary tree architecture for multi-class classification. Neural Inf Proc—Lett Rev 2(3)
15. Jing C, Sha J. An efficient implementation of 2D convolution in CNN. In: IEICE electronics express, vol 14. pp. 20161134–20161134. https://doi.org/10.1587/elex.13.20161134
16. Vetterli M, Herley C (Sept 1992) Wavelets and filter banks: theory and design. IEEE Trans Signal Process 40(9):2207–2232. https://doi.org/10.1109/78.157221
17. Fayek HM, Lech M, Cavedon L (2015) Towards real-time speech emotion recognition using deep neural networks 2015. In: 9th international conference on signal processing and communication systems (ICSPCS). pp 1–5. https://doi.org/10.1109/ICSPCS.2015.7391796